Similarity & Link Analysis

Stony Brook University CSE545, Fall 2016

Finding Similar "Items"



(http://blog.soton.ac.uk/hive/2012/05/10/r ecommendation-system-of-hive/)







(http://www.datacommunitydc.org/blog/20 13/08/entity-resolution-for-big-data)

Finding Similar "Items": What we will cover

- Shingling
- Minhashing
- Locality-sensitive hashing
- Distance Metrics

Document Similarity

Challenge: How to represent the document in a way that can be efficiently encoded and compared?

Goal: Convert documents to sets



Goal: Convert documents to sets



k-shingles (aka "character n-grams")- sequence of k characters

E.g. *k*=2 doc="abcdabd" singles(doc, 2) = {ab, bc, cd, da, bd}

Goal: Convert documents to sets



k-shingles (aka "character n-grams")- sequence of k characters

- E.g. *k*=2 doc="abcdabd" singles(doc, 2) = {ab, bc, cd, da, bd}
- Similar documents have many common shingles
- Changing words or order has minimal effect.
- In practice use 5 < k < 10

Goal: Convert documents to sets



Large enough that any given shingle appearing a document is highly unlikely (e.g. < .1% chance)

Can hash large shingles to smaller (e.g. 9-shingles into 4 bytes)

Can also use words (aka n-grams).

acters

a, bd}

- Similar documents have many common shingles
 Changing will fis or order has minimal effect.
- In practice use 5 < k < 10

Problem: Even if hashing, sets of shingles are large (e.g. 4 bytes => 4x the size of the document).

Goal: Convert sets to shorter ids, signatures

Goal: Convert sets to shorter ids, signatures

Ghara	CLEIIS		aur,	Λ .	
Element	S_1	S_2	S_3	S_4	
a	1	0	0	1	
b	0	0	1	0	
c	0	1	0	1	
d	1	0	1	1	
e	0	0	1	0	

Characteristic Matrix V:

(Leskovec at al., 2014; http://www.mmds.org/)

often very sparse! (lots of zeros)

Jaccard Similarity:



Characteristic Matrix:

	<i>S</i> ₁	<i>S</i> ₂
ab	1	1
bc	0	1
de	1	0
ah	1	1
ha	0	0
ed	1	1
ca	0	1

Jaccard Similarity: $sim(S_1, S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2}$

Characteristic Matrix:

	<i>S</i> ₁	<i>S</i> ₂	
ab	1	1	* *
bc	0	1	*
de	1	0	*
ah	1	1	**
ha	0	0	
ed	1	1	**
са	0	1	*

Jaccard Similarity: $sim(S_1, S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2}$

Characteristic Matrix:

	<i>S</i> ₁	<i>S</i> ₂	
ab	1	1	* *
bc	0	1	*
de	1	0	*
ah	1	1	**
ha	0	0	
ed	1	1	**
са	0	1	*

Jaccard Similarity: $sim(S_1, S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2}$

 $sim(S_1, S_2) = 3 / 6$ # both have / # at least one has

Problem: Even if hashing, sets of shingles are large (e.g. 4 bytes => 4x the size of the document).

Characteristic Matrix: X

	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
ab	1	0	1	0
bc	1	0	0	1
de	0	1	0	1
ah	0	1	0	1
ha	0	1	0	1
ed	1	0	1	0
са	1	0	1	0

Approximate Approach:

1) Instead of keeping whole characteristic matrix, just keep first row where 1 is encountered.

2) Shuffle and repeat to get a "signature" for each set.



Characteristic Matrix:

	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
ab	1	0	1	0
bc	1	0	0	1
de	0	1	0	1
ah	0	1	0	1
ha	0	1	0	1
ed	1	0	1	0
са	1	0	1	0

Minhash function: h

 Based on permutation of rows in the characteristic matrix, *h* maps sets to first row where set appears.

Characteristic Matrix:

	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
ab	1	0	1	0
bc	1	0	0	1
de	0	1	0	1
ah	0	1	0	1
ha	0	1	0	1
ed	1	0	1	0
са	1	0	1	0

permuted order 1 ha 2 ed 3 ab 4 bc 5 ca 6 ah 7 de

Minhash function: *h*

 Based on permutation of rows in the characteristic matrix, *h* maps sets to first row where set appears.

Characteristic Matrix:

		<i>S</i> ₁	<i>S</i> ₂	S ₃	<i>S</i> ₄
3	ab	1	0	1	0
4	bc	1	0	0	1
7	de	0	1	0	1
6	ah	0	1	0	1
1	ha	0	1	0	1
2	ed	1	0	1	0
5	са	1	0	1	0

permuted order 1 ha 2 ed 3 ab 4 bc 5 ca 6 ah 7 de

Minhash function: *h*

 Based on permutation of rows in the characteristic matrix, *h* maps sets to first row where set appears.

Characteristic Matrix:

		<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
3	ab	1	0	1	0
4	bc	1	0	0	1
7	de	0	1	0	1
6	ah	0	1	0	1
1	ha	0	1	0	1
2	ed	1	0	1	0
5	са	1	0	1	0

permuted order 1 ha 2 ed 3 ab 4 bc 5 ca 6 ah 7 de

Minhash function: *h*

 Based on permutation of rows in the characteristic matrix, *h* maps sets to first row where set appears.

> $h(S_1) = ed #permuted row 2$ $h(S_2) = ha #permuted row 1$ $h(S_3) =$

Characteristic Matrix:

		<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
3	ab	1	0	1	0
4	bc	1	0	0	1
7	de	0	1	0	1
6	ah	0	1	0	1
1	ha	0	1	0	1
2	ed	1	0	1	0
5	са	1	0	1	0

permuted order 1 ha 2 ed 3 ab 4 bc 5 ca 6 ah 7 de

(Leskovec at al., 2014; http://www.mmds.org/)

Minhash function: *h*

 Based on permutation of rows in the characteristic matrix, *h* maps sets to first row where set appears.

> $h(S_1) = ed #permuted row 2$ $h(S_2) = ha #permuted row 1$ $h(S_3) = ed #permuted row 2$ $h(S_4) =$

Characteristic Matrix:

		<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
3	ab	1	0	1	0
4	bc	1	0	0	1
7	de	0	1	0	1
6	ah	0	1	0	1
1	ha	0	1	0	1
2	ed	1	0	1	0
5	са	1	0	1	0

permuted order 1 ha 2 ed 3 ab 4 bc 5 ca 6 ah 7 de

(Leskovec at al., 2014; http://www.mmds.org/)

Minhash function: *h*

 Based on permutation of rows in the characteristic matrix, *h* maps sets to first row where set appears.

> $h(S_1) = ed$ #permuted row 2 $h(S_2) = ha$ #permuted row 1 $h(S_3) = ed$ #permuted row 2 $h(S_4) = ha$ #permuted row 1

Characteristic Matrix:

		<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
3	ab	1	0	1	0
4	bc	1	0	0	1
7	de	0	1	0	1
6	ah	0	1	0	1
1	ha	0	1	0	1
2	ed	1	0	1	0
5	са	1	0	1	0

(Leskovec at al., 2014; http://www.mmds.org/)

Minhash function: h

1

• Based on permutation of rows in the characteristic matrix, *h* maps sets to rows.

Signature matrix: M

 Record first row where each set had a 1 in the given permutation

$$\begin{array}{|c|c|c|c|c|c|c|c|}\hline & S_1 & S_2 & S_2 & S_3 & S_4 \\ \hline & h_1 & 2 & 1 & 2 & 1 \\ \hline \end{array}$$

 $h_1(S_1) = ed \text{ #permuted row}$ $h_1(S_2) = ha \text{ #permuted row}$ $h_1(S_2) = ed \text{ #permuted row}$

Characteristic Matrix:

		<i>S</i> ₁	<i>S</i> ₂	S ₃	<i>S</i> ₄
3	ab	1	0	1	0
4	bc	1	0	0	1
7	de	0	1	0	1
6	ah	0	1	0	1
1	ha	0	1	0	1
2	ed	1	0	1	0
5	са	1	0	1	0

(Leskovec at al., 2014; http://www.mmds.org/)

Minhash function: h

• Based on permutation of rows in the characteristic matrix, *h* maps sets to rows.

Signature matrix: M

2

1

 Record first row where each set had a 1 in the given permutation

 $h_1(S_1) = ed #permuted row$

 $h_1(S_2) = ha \# permuted row$

h(S) = ed # permuted row

Characteristic Matrix:

		<i>S</i> ₁	<i>S</i> ₂	S ₃	<i>S</i> ₄
3	ab	1	0	1	0
4	bc	1	0	0	1
7	de	0	1	0	1
6	ah	0	1	0	1
1	ha	0	1	0	1
2	ed	1	0	1	0
5	са	1	0	1	0

Minhash function: h

• Based on permutation of rows in the characteristic matrix, *h* maps sets to rows.

Signature matrix: M

1

 Record first row where each set had a 1 in the given permutation

$$\begin{array}{|c|c|c|c|c|c|c|c|} S_1 & S_2 & S_3 & S_4 \\ \hline & & & & \\ h_1 & 2 & 1 & 2 & 1 \\ \end{array}$$

$$h_1(S_1) = ed #permuted row$$

 $h_1(S_2) = ha \# permuted row$

h(S) = ed #permuted row

Characteristic Matrix:

		<i>S</i> ₁	S_2	S_{3}	<i>S</i> ₄
3	ab	1	0	1	0
4	bc	1	0	0	1
7	de	0	1	0	1
6	ah	0	1	0	1
1	ha	0	1	0	1
2	ed	1	0	1	0
5	са	1	0	1	0
	3 4 7 6 1 2 5	3 ab 4 bc 7 de 6 ah 1 ha 2 ed 5 ca	Image: style	Image: series of the series	S_1 S_2 S_3 3ab1014bc1007de0106ah0101ha0102ed1015ca101

Minhash function: h

• Based on permutation of rows in the characteristic matrix, *h* maps sets to rows.

Signature matrix: M

 Record first row where each set had a 1 in the given permutation

	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
h_1	2	1	2	1
<i>h</i> ₂				

Characteristic Matrix:

		<i>S</i> ₁	S_2	S_{3}	<i>S</i> ₄
3	ab	1	0	1	0
4	bc	1	0	0	1
7	de	0	1	0	1
6	ah	0	1	0	1
1	ha	0	1	0	1
2	ed	1	0	1	0
5	са	1	0	1	0
	3 4 7 6 1 2 5	3 ab 4 bc 7 de 6 ah 1 ha 2 ed 5 ca	Image: style	Image: series of the series	S_1 S_2 S_3 3ab1014bc1007de0106ah0101ha0102ed1015ca101

Minhash function: *h*

• Based on permutation of rows in the characteristic matrix, *h* maps sets to rows.

Signature matrix: M

 Record first row where each set had a 1 in the given permutation

	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
h ₁	2	1	2	1
<i>h</i> ₂	2	1	4	1

Characteristic Matrix:

				<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
1	4	3	ab	1	0	1	0
3	2	4	bc	1	0	0	1
7	1	7	de	0	1	0	1
6	3	6	ah	0	1	0	1
2	6	1	ha	0	1	0	1
5	7	2	ed	1	0	1	0
4	5	5	са	1	0	1	0

Minhash function: h

• Based on permutation of rows in the characteristic matrix, *h* maps sets to rows.

Signature matrix: M

 Record first row where each set had a 1 in the given permutation

	<i>S</i> ₁	<i>S</i> ₂	S ₃	<i>S</i> ₄
h_1	2	1	2	1
h ₂	2	1	4	1
h ₃				

Characteristic Matrix:

				<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
1	4	3	ab	1	0	1	0
3	2	4	bc	1	0	0	1
7	1	7	de	0	1	0	1
6	3	6	ah	0	1	0	1
2	6	1	ha	0	1	0	1
5	7	2	ed	1	0	1	0
4	5	5	са	1	0	1	0

Minhash function: h

• Based on permutation of rows in the characteristic matrix, *h* maps sets to rows.

Signature matrix: M

 Record first row where each set had a 1 in the given permutation

	<i>S</i> ₁	<i>S</i> ₂	S ₃	<i>S</i> ₄
h ₁	2	1	2	1
<i>h</i> ₂	2	1	4	1
h ₃	1	2	1	2

Characteristic Matrix: X

				<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
1	4	3	ab	1	0	1	0
3	2	4	bc	1	0	0	1
7	1	7	de	0	1	0	1
6	3	6	ah	0	1	0	1
2	6	1	ha	0	1	0	1
5	7	2	ed	1	0	1	0
4	5	5	са	1	0	1	0

Minhash function: h

• Based on permutation of rows in the characteristic matrix, *h* maps sets to rows.

Signature matrix: M

• Record first row where each set had a 1 in the given permutation

	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
h ₁	2	1	2	1
h ₂	2	1	4	1
h ₃	1	2	1	2

Characteristic Matrix:

Property of signature matrix: The probability for any h_i (i.e. any row), that $h_i(S_1) = h_i(S_2)$ is the same as $Sim(S_1, S_2)$

				<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
1	4	3	ab	1	0	1	0
3	2	4	bc	1	0	0	1
7	1	7	de	0	1	0	1
6	3	6	ah	0	1	0	1
2	6	1	ha	0	1	0	1
5	7	2	ed	1	0	1	0
4	5	5	са	1	0	1	0

	<i>S</i> ₁	<i>S</i> ₂	S ₃	<i>S</i> ₄
h ₁	2	1	2	1
h ₂	2	1	4	1
h ₃	1	2	1	2

Characteristic Matrix:

				<i>S</i> ₁	<i>S</i> ₂	S ₃	<i>S</i> ₄
1	4	3	ab	1	0	1	0
3	2	4	bc	1	0	0	1
7	1	7	de	0	1	0	1
6	3	6	ah	0	1	0	1
2	6	1	ha	0	1	0	1
5	7	2	ed	1	0	1	0
4	5	5	са	1	0	1	0

Property of signature matrix: The probability for any h_i (i.e. any row), that $h_i(S_1) = h_i(S_2)$ is the same as $Sim(S_1, S_2)$

Thus, similarity of signatures S_1 , S_2 is the fraction of minhash functions (i.e. rows) in which they agree.

	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
h ₁	2	1	2	1
h ₂	2	1	4	1
h ₃	1	2	1	2





Characteristic Matrix:

				<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
1	4	3	ab	1	0	1	0
3	2	4	bc	1	0	0	1
7	1	7	de	0	1	0	1
6	3	6	ah	0	1	0	1
2	6	1	ha	0	1	0	1
5	7	2	ed	1	0	1	0
4	5	5	са	1	0	1	0

Property of signature matrix: The probability for any h_i (i.e. any row), that $h_i(S_1) = h_i(S_2)$ is the same as $Sim(S_1, S_2)$

Thus, similarity of signatures $S_{1^{\prime}}$, S_{2} is the fraction of minhash functions (i.e. rows) in which they agree.

	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
h ₁	2	1	2	1
h ₂	2	1	4	1
h ₃	1	2	1	2

Estimated Sim(S₁, S₃) = agree / all = 2/3

Real Sim(S₁, S₃) = Type a / (a + b + c) = 3/4

Characteristic Matrix:

				<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
1	4	3	ab	1	0	1	0
3	2	4	bc	1	0	0	1
7	1	7	de	0	1	0	1
6	3	6	ah	0	1	0	1
2	6	1	ha	0	1	0	1
5	7	2	ed	1	0	1	0
4	5	5	са	1	0	1	0

Property of signature matrix: The probability for any h_i (i.e. any row), that $h_i(S_1) = h_i(S_2)$ is the same as $Sim(S_1, S_2)$

Thus, similarity of signatures S_1 , S_2 is the fraction of minhash functions (i.e. rows) in which they agree.

	<i>S</i> ₁	<i>S</i> ₂	<i>S</i> ₃	<i>S</i> ₄
h ₁	2	1	2	1
<i>h</i> ₂	2	1	4	1
h ₃	1	2	1	2

Estimated Sim(S₁, S₃) = agree / all = 2/3

Real Sim(S₁, S₃) = Type a / (a + b + c) = 3/4

Try Sim(S $_2$, S $_4$) and Sim(S $_1$, S $_2$)
In Practice

Problem:

- Can't reasonably do permutations (huge space)
- Can't randomly grab rows according to an order (random disk seeks = slow!)

In Practice

Problem:

- Can't reasonably do permutations (huge space)
- Can't randomly grab rows according to an order (random disk seeks = slow!)

Solution: Use "random" hash functions.

- Setup:
 - Pick ~100 hash functions, hashes
 - Store M[i][s] = a potential minimum h_i(r)
 #initialized to infinity (num hashs x num sets)

Solution: Use "random" hash functions.

- Setup:
 - Pick ~100 hash functions, hashes
 - Store M[i][s] = a potential minimum h_i(r)
 #initialized to infinity (num hashs x num sets)

• Algorithm:

for r in rows of cm: #cm is characteristic matrix compute $h_i(r)$ for all i in hashes #precompute 100 values for each set s in row r:

if cm[r][s] == 1:

for i in hashes: #check which hash produces smallest value
 if h_i(r) < M[i][s]: M[i][s] = h_i(r)

Algorithm: –

Solution: Use "random" hash functions.

- Setup:
 - Pick ~100 hash functions, hashes
 - Store M[i][s] = a potential minimum $h_i(r)$ #initialized to i

Known as "efficient minhashing".

for r in rows of cm: #cm is characteristic matrix
 compute h_i(r) for all i in hashes #precompute 100 values
 for each set s in row r:

if cm[r][s] == 1:

for i in hashes: #check which hash produces smallest value
 if h_i(r) < M[i][s]: M[i][s] = h_i(r)

What hash functions to use?

Start with 2 decent hash functions

e.g. $h_a(x) = ascii(string) \% large_prime_number$ $h_b(x) = (3*ascii(string) + 16) \% large_prime_number$

https://www.eecs.harvard.edu/~michaelm/postscripts/rsa2008.pdf

What hash functions to use?

Start with 2 decent hash functions

e.g. $h_a(x) = ascii(string) \% large_prime_number$ $h_b(x) = (3*ascii(string) + 16) \% large_prime_number$

Add together multiplying the second times i:

 $h_i(x) = h_a(x) + i^*h_b(x)$ e.g. $h_5(x) = h_a(x) + 5^*h_b(x)$

https://www.eecs.harvard.edu/~michaelm/postscripts/rsa2008.pdf

Problem: Even if hashing, sets of shingles are large (e.g. 4 bytes => 4x the size of the document).

Problem: Even if hashing, sets of shingles are large (e.g. 4 bytes => 4x the size of the document).

New Problem: Even if the size of signatures are small, it can be computationally expensive to find similar pairs.

E.g. 1m documents; 1,000,000 choose 2 = 500,000,000,000 pairs

Goal: find pairs of minhashes likely to be similar (in order to then test more precisely for similarity).

Candidate pairs: pairs of elements to be evaluated for similarity.

Goal: find pairs of minhashes likely to be similar (in order to then test more precisely for similarity).

Candidate pairs: pairs of elements to be evaluated for similarity.

If we wanted the similarity for all pairs of documents, could anything be done?

Goal: find pairs of minhashes likely to be similar (in order to then test more precisely for similarity).

Candidate pairs: pairs of elements to be evaluated for similarity.

Approach: Hash multiple times over subsets of data: similar items are likely in the same bucket once.

Goal: find pairs of minhashes likely to be similar (in order to then test more precisely for similarity).

Candidate pairs: pairs of elements to be evaluated for similarity.

Approach: Hash multiple times over subsets of data: similar items are likely in the same bucket once.

Approach from MinHash: Hash columns of signature matrix

→ Candidate pairs end up in the same bucket.

Step 1: Add bands

Locality-Sensitive Hashing



(Leskovec at al., 2014; http://www.mmds.org/)

Step 1: Add bands

Locality-Sensitive Hashing



(Leskovec at al., 2014; http://www.mmds.org/)

Step 1: Add bands Step 2: Hash columns within bands



(Leskovec at al., 2014; http://www.mmds.org/)



Step 1: Add bands Step 2: Hash columns within bands

(Leskovec at al., 2014; http://www.mmds.org/)



Step 1: Add bands Step 2: Hash columns within bands

(Leskovec at al., 2014; http://www.mmds.org/)





Step 1: Add bands Step 2: Hash columns within bands

Document Similarity Pipeline



- 100,000 documents
- 100 random permutations/hash functions/rows
 => if 4byte integers then 40Mb to hold signature matrix
 => still 100k choose 2 is a lot (~5billion)

- 100,000 documents
- 100 random permutations/hash functions/rows
 => if 4byte integers then 40Mb to hold signature matrix
 => still 100k choose 2 is a lot (~5billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity ; for any row $p(S_1 == S_2) = .8$

- 100,000 documents
- 100 random permutations/hash functions/rows
 => if 4byte integers then 40Mb to hold signature matrix
 => still 100k choose 2 is a lot (~5billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity ; for any row $p(S_1 == S_2) = .8$

 $P(S_1 = S_2 | b)$: probability S1 and S2 agree within a given band

- 100,000 documents
- 100 random permutations/hash functions/rows
 => if 4byte integers then 40Mb to hold signature matrix
 => still 100k choose 2 is a lot (~5billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity ; for any row $p(S_1 == S_2) = .8$

$$\begin{split} \mathsf{P}(\mathsf{S}_1 == \mathsf{S}_2 \mid \mathsf{b}): \text{ probability S1 and S2 agree within a given band} \\ &= 0.8^5 = .328 \quad = > \quad \mathsf{P}(\mathsf{S}_1 != \mathsf{S}_2 \mid \mathsf{b}) = 1 - .328 = .672 \\ \mathsf{P}(\mathsf{S}_1 != \mathsf{S}_2): \text{ probability S1 and S2 do not agree in any band} \end{split}$$

- 100,000 documents
- 100 random permutations/hash functions/rows
 => if 4byte integers then 40Mb to hold signature matrix
 => still 100k choose 2 is a lot (~5billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity ; for any row $p(S_1 == S_2) = .8$

$$\begin{split} \mathsf{P}(\mathsf{S}_1 == \mathsf{S}_2 \mid \mathsf{b}): \text{ probability S1 and S2 agree within a given band} \\ &= 0.8^5 = .328 \quad => \quad \mathsf{P}(\mathsf{S}_1 != \mathsf{S}_2 \mid \mathsf{b}) = 1 - .328 = .672 \\ \mathsf{P}(\mathsf{S}_1 != \mathsf{S}_2): \text{ probability S1 and S2 do not agree in any band} \\ &= .672^{20} = .00035 \end{split}$$

- 100,000 documents
- 100 random permutations/hash functions/rows
 => if 4byte integers then 40Mb to hold signature matrix
 => still 100k choose 2 is a lot (~5billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity ; for any row $p(S_1 == S_2) = .8$

 $P(S_1 == S_2 | b): \text{ probability S1 and S2 agree within a given band}$ = 0.8⁵ = .328 => $P(S_1 != S_2 | b) = 1..328 = .672$ $P(S_1 != S_2): \text{ probability S1 and S2 do not agree in any band}$ = .672²⁰ = .00035

What if wanting 40% Jaccard Similarity?

Pipeline gives us a way to find *near-neighbors* in *high-dimensional space* based on Jaccard Distance (1 - Jaccard Sim).



(http://rosalind.info/glossary/euclidean-distance/)

Pipeline gives us a way to find *near-neighbors* in *high-dimensional space* based on Jaccard Distance (1 - Jaccard Sim).

Typical properties of a distance metric, *d*:

d(x, x) = 0d(x, y) = d(y, x)

 $d(x, y) \le d(x,z) + d(z,y)$



(http://rosalind.info/glossary/euclidean-distance/)

Pipeline gives us a way to find *near-neighbors* in *high-dimensional space* based on Jaccard Distance (1 - Jaccard Sim).

There are other metrics of similarity. e.g:

- Euclidean Distance
- Cosine Distance

...

- Edit Distance
- Hamming Distance

Pipeline gives us a way to find *near-neighbors* in *high-dimensional space* based on Jaccard Distance (1 - Jaccard Sim).

There are other metrics of similarity. e.g:

• Euclidean Distance

$$distance(X,Y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2} \quad (``L2 Norm")$$

n

• Cosine Distance

Edit Distance

. . .

Hamming Distance

Pipeline gives us a way to find *near-neighbors* in *high-dimensional space* based on Jaccard Distance (1 - Jaccard Sim).

There are other metrics of similarity. e.g:

- Euclidean Distance
- Cosine Distance

• Edit Distance

. . .

Hamming Distance



Locality Sensitive Hashing - Theory

LSH Can be generalized to many distance metrics by converting output to a probability and providing a lower bound on probability of being similar.

Locality Sensitive Hashing - Theory

LSH Can be generalized to many distance metrics by converting output to a probability and providing a lower bound on probability of being similar.

- E.g. for euclidean distance:
- Choose random lines (analogous to hash functions in minhashing)
- Project the two points onto each line; match if two points within an interval

Link Analysis

The Web, circa 1998

	ALTAVISTA View Multimedia From Our Vantage Point
AUTORITEL	Car Buying & Car Insurance Pain Relief
Click he Search th	re for advertising information - reach millions every month!
Search with	Submit Digital's Alta Vista [Advanced Search] [Add URL]
Ma	Contests Creative Web



•Xeite	search reviews city.net I	ives tours
DIRECT	Excite Search: twice the power of the comp What:	actition.
Turbo Search!" Download Excite Direct	Where: World Wide Web	[Help] [Advanced Search]
Take an	INTEGRATED BROWSING, EMAIL, NEWSGROUPS AND PAGE CREATION.	Never Never
ExciteSeeing Tour	Excite Reviews: site reviews by the web's	best editorial team.



tares' create your own webcam	Click to develoed Takes! Messenger	free from anywhere
	Search ad	vanced search
Yt Shopping Depts: Book	. CDs. Computers, DVDs Stores (Sap. Clinique. Coach and more
Auctions Autos Classifieds Sho et Carers Chat Clubs GeoCities and Arish Book Briefrang Calmi	opping Transf Yellon Ppi Mapa Modia Greetings Mail Members Messenger at My Yahoo! PerDirect For Games 1	Einance Quotes Mens Sparts We Mable - Personals - People Search - lids - Mexico - Muns - Ratio - TV - m
Yahoo! Auctions	Bid burn or sell anything!	In the News
Categories	ltems	· U.S. missite SIM apr plane tab
Antiques Computers	Xeea Dale Eambardt	- Sources: Condit adouts to nemal
Caunas Electronica	Golf Clubs Beatles Yesterday	relationship with missing intern
Camic Books Stamps	Extension Second	 Attemny Barry Levin frund dead
Bateball Cards - McGaire A.Ro	4 Jeter Bonds, Sosa, Geffey Jr., Johan	* Dale Earnhardt Jr. witte Pepte 400
AND DESCRIPTION OF A DE	e com avenue voice seengato hulli	• Wimbledon - Tour de France
Arts & Humanities	News & Media	E.O.S.
Literature, Photography -	Full Conwage, Newspapers, TV-	Marketplace
	State State State State State	- new Britney shaps London.
Business & Economy	Recreation & Sports	- 11 Store - become part of Yahao!
B2B Emance Shepping John-	Sparts Travel Astes Outdesta-	Shepping
Computers 8 Internet	Deferment	- 11 Ceretra - find a job, pest your
Computers a internet	Kelerence	Descane
BRIDE, N.Y. S. SECOND. SHEET-	COMPANY, DESIGNATED CONSIDERE-	accessories
Education	Regional	Proadcast Events
College and University, K-12	Countries, Regions, US States	- Jose FT - PGA Wasteen Ones
		blied, 102, Arrist of the worth
Entertainment	Science	INCOME AND ADDRESS OF
CostLinks Menter Burner Mann	Annali Amonosy Inpassing-	Inside Yahool
Government	Social Science	- 17 Games - backgammon, eaches,
Electrons, Military, Law, Taxes -	Archaeology, Economica, Languages-	hearts, chess, pinochle
Construction of the second	and the second second second	- 17 Meyics - Scary Movie 2. Kinn
Health	Society & Culture	of the Dragon, Cats and Dogs
Medicine Diseases Drugs Etness	People, Environment, Religion	midnesson venian
Pow	and by	· 11 Photos - post your party pice
	IMPIAO.	
Local Yahoots	annen beb Versen Seen Brader I	W. & Instand
Arte Perific : Aria - Australia	& NZ - China - MK - India - Japan - Korea-	Sectore - Taiwat
American Accention Board	Canada - Chinese - Meting - Spanish	and the second
U.S. Catter : Adapta - Beston -	Chicago - Dulles FW - LA - NYC - SF Bay-	Wath DC - mare-
Mare Yaboo's		
Guider : Antes - Darr John -	Catents - Health - Living - Ostdoors - Pets -	Real Lotate - Tabackgana!
Entertablement : Astrology - B	realizat - Events - Ganes - Morries - Massi	- Radio - Tickets - TV - more
Pononce : Banking - Bill Pay -	nourmos - Leans - Lates - Futurice vision -	2000
CONTRACT DESCRIPTION - DOCEMENTS -	COMPRESSION - NAMESARD - TRIOW PART	2 - EXCER

Make Tabeol your home page

Haw to Suggest a Site - Company Info - Copyright Policy - Terms of Service - Contributors - Jaba - Advertising

Copyright @ 2001 Yahoo! Inc. All rights reserved. Privacy Policy

The Web , circa 1998

AUTO STEL	Car Buying & Car Insurance Pain Relief
y and insure new cars & tru	cks online
Click he	re for advertising information - reach millions every month!
Search th	e Web wand Display the Results in Standard Form w



Match keywords, language (*information retrieval*) Explore directory

excite	people finder maps yellow	live tours
"Turbo Search!" Download Excite Direct	What: Where: World Wide Web	Help [Help] [Advanced Search]
Take an ExciteSeeing Tour	INTEGRATED BROWSING, EMAIL, NEWSCROUPS AND PACE CREATION Excite Reviews: site reviews by the web"	s best editorial team.



Y1 Shopping Depts: Books, CDs, Computers, DVDs Stores, Gap, Chright, Coach a Austrom Anna Gausfield, Shopping Tarni, Talkoz, Pay Maya Mella Finance Queres, Nara, Chas, Geordian, Mail Manten, Massanar, Malathe Finance, Queres, Mar, Sanapar, Analathe, Maximum Malather, Fard Analather, Markabara, Chas, Bachall Cach, Martin, A.Rod Jack, Book, Sana, Chas, Bachall Cach, Martin, A.Rod Jack, Book, Sana, Sanapar, Charasher, Sanapar, Chas, Sanapar, Chas, Sanapar, Charasher, Sanapar, Sanap	pping Depts: Books, CDs.		and the second se
Antonne Anne Chanfiels Stepping Torrel Villez-Epi Mapi Molis Emanes Queres Near - Carren Char China Sections Sections Mell Medices Measures Measure Medic Tenness Create Anton Mathema Measures Stepping Torrel Villez-Epi Mapi Molis Emanes Queres Near - Walk Add Book Sendias Sections Sections Mellin Medices Measure Medica Mathema Measures Compared Provided China Section Sections Sections Chart Medica Medica Section Sections Section Measures Constructions Section Measures Constructions Section Sections Section Section Section Section Section Sections Measures Construction Section Sectio		Computers, DVDs Stores (Gap. Clinique, Coach and more
Yahool Auctions Bit key, or sell arythong? ategories Saman Saman <th>Autor Classifieds Shopping Chat Chiles GeoCities Greetin Book Briefcase Calendar My</th> <th>Interel Velice Pgr Magis Media gs Meil - Menters - Messenger Jahool - ParDirect Fan Games - J</th> <th>Eleance Quotes Merce Spatte Was Mable - Freionals - People Search - B Kida - Mavies - Masie - Rache - IV - me</th>	Autor Classifieds Shopping Chat Chiles GeoCities Greetin Book Briefcase Calendar My	Interel Velice Pgr Magis Media gs Meil - Menters - Messenger Jahool - ParDirect Fan Games - J	Eleance Quotes Merce Spatte Was Mable - Freionals - People Search - B Kida - Mavies - Masie - Rache - IV - me
ategories Cangatan Sangari Cangatan Santani Electronics Santani Electronics Santani Electronics Santani Electronics Santani Electronics Santani Electronics Barchall Cade Molform Barchall Cade Molform Santani Electronics Barchall Cade Molform Attance Barchall Cade Internes Santani Santani Attance Internes Barchall Cade Internet Barchall Cade Internet Barchal	Yahoo! Auctions . Bid to	zv. or sell avvrbing!	In the News
Sandarsi Comparent Name Disk Landardi Sandarsi Electronical Golf Chas Senters Vestorday Sandarsi Spars Could Sold Chas Senters Vestorday Sandarsi Standarsi Senters Vestorday Sandarsi Standarsi Senters Vestorday Sandarsi Standarsi Senters Vestorday Sandarsi Standarsi Senters Vestorday Sandarsi News & Modia Senters Vestorday Sandarsi News & Modia Senters Vestorday Sandarsi Senters Senters Senters Senters Santaria Senters Senters Senters Santaria Senters Senters Senters Santaria Senters <th>ltem</th> <th></th> <th>+ U.S. miecta \$154 apy plane tab</th>	ltem		+ U.S. miecta \$154 apy plane tab
Interna - Electronica - Golf Chila - Sendra Yestendov Interna - Electronica - Golf Chila - Sendra Yestendov International - Margin - Sendra - Sendra Yestendov Electronical - Margine - Sendra - Sendra - Yestendov International - Margine - Sendra - Send	- Computers - Xee	Dale Earthardt	- Sources: Condit admits to sexual
International Sponta Cardin and Contents Anterneo Renu Levini Basic Baok Stange - Debrano Contents Data East Contents Data East Contents Baschall Cards - McGrain - Allod Jeen Bonds Sons Geffey In, Linter Data Easthell Contents Data Easthell Contents Its & Humanities Easthell Cards - McGrain - Allod Jeen Bonds Sons Geffey In, Linter Data Easthell Contents Markedow - Tendel Sons Geffey In, Linter Its & Humanities Easthell Connege Strangagers, TE- Ball Connege Strangagers, TE- Wintlebox - Tendel Sons Geffey In, Linter Based Cards - McGrain A Sports Easthell Connege Strangagers, TE- Based Connege Strangagers, TE- Wintlebox - Ferrange 200 Based Cards - Strangers, Schu- Based Connege Strangagers, TE- Wintlebox - Tendel Strangers, Tender, Tender Strangers, Tender, Tender Strangers, Tender,	Electronics Goff	Clubs Seatles Yesterday	relationship with missing inten
And Casis - Material And Juer Books Sone Gentry Julian Barchall Casis - Material And Juer Books Sone Gentry Julian Sone Sone Sone Sone Sone Sone Sone S	Sports Cards Bis	Langabargar	· Attempy Barry Levin frund dead
Extended Looks - Soferen Autor Annue Roman Social Gentry Jr. Humen Its A Humanities Instance, Thotography. Its Tances, Stepson, Janua, Its Stepson, Ste	ka Stamps Deb	1820 Sce.com	· Date Earnhardt Jr. winn Pepra 400
Instanticies News & Media teames, Biotography Ball Cornege, Neuropart, Tri- tauiness & Economy Recreation & Sports Ball Cornege, Neuropart, Tri- Recreation & Sports computers & Internet Recreation & Contributions ducation Recional ducation Connet, Repres, US Suites discent University, Rep. La Science antertainment Science actions, Statuy, Lee, Tauss Action Actionation, US Suites sovernment Social Science Anterial Science Anterial Actionation, Languagen action Statuy, Lee, Tauss Social Science Anterials, Despine, Drugs, France Produc Environment, Region, Languagen sovernment Sociely & Culture earth Produc Environment Relignon action Statuy, Lee, Tauss Produc Environment Relignon Control, Contro	Lands - Distantine, Autori, 1898, 1	ionas, sosa, testey it., ittera	+ Wimbledon - Tour de France
Instance News & Media Lattornage, Neurogens, TT- Baltornage, Neurogens, TT- usiness & Economy Baltornage, Neurogens, TT- in Jamme, Shepping, John- Bernit, Janual, Astra, Outlann- computers & Internet Reference timmt, WWW, Software, Same- Litrans, Dictionation, Constition- ducation Reference timmt, WWW, Software, Same- Litrans, Dictionation, Constition- ducation Reference timmt, WWW, Software, Same- Councils, Region M ducation Science atlack, Morean, Banco, Manie Science Social Science Antaniogre, Economic, Languagen, Science,	1000 C 1000	and the second se	Reff.
Indiana, Diseases, Drugs, Errors, Drugs, Discover, Dennes, Marine, Diseases, Drugs, Errors, Status, Castanon, Marine, Mar	umanibes New	s & Media	Harketplace
Justiness & Economy Recreation & Sports B. Tamos, Stepsing, John- Sparts, Tamos, Auto, Outdeen- Omputers & Internet Reference Librares, Dational Recional Garanto, Marco Regional Garanto, Marco Science Antano Autonomy, Lagnering- Social Science Antano Autonomy, Lagnering- Social Science Antanology, Lenson, Drugs, France Antanology, Lenson, Languagen- Resting, Drugs, France Social Science Antanology, Lenson, Drugs, France Social Science Antanology, Lenson, Drugs, France Proto, Environment, Religen- Barbin, Drugs, France Proto, Environment, Religen- Commer, Strange, Lenson, Drugs, France, Strange, Strange	notography - Edit C	analy seals and seals	- new Britney, shops London;
Annuel, Margania, Johnson, James Jana, Janie Aura, Carlasen, Janier, Janie Janie, Janie Janie, Carlasen, Janier, Janieren, Janierev, Janierev, Janier, Janier, Janier, Janier, Janier, Ja	s & Economy Rec	reation & Sports	Epinode 4 - spansared by Papai
Analysis Analysis Analysis Reference Analysis Reference Analysis Analysis Analysis Regional Connict, Repairs, UK Stata Science Analysis Analysis Analysis Science Analysis Analysis Socience Analysis Analysis Analysis Socience Analysis Analysis Analysis Socience Analysis Analysis Analysis Backers Analysis Backers Socience Analysis Analysis Backers Analysis Backers Socience Analysis Analysis Backers Backers	s Sherrane John South	Travel Actes Outdoors	- 11 Store - become part of Yahoo!
Someware Sames Reference ducation Internets Divisioning, Contations, ducation - Internets Divisioning, Contations, ducation ducation Regional - Internets Divisioning, Contations, ducation ducation Regional - Internets Divisioning, Contations, ducation ducation Science - Internet Division file astilarda Microsofty, Law, Taxon Science - Internet Division file inverse Social Science - Internet Division ducation Social Science - Internet Division internet Science - Annuals Antonomy, Imparentage - Internet Division file internet Minisco Line, Line, Internet Science - Internet Science - Internet Division internet Minisco Line, Division of the Divisio	- mapping that - man	The second second	Shopping
terrent, WWW, Software, Games -, Literates, Distinguist, Constitions -, Kalada phages, series december of the Construction -, Kalada phages, series december	ers & Internet Refe	rence	resume
Social Science Social Science Instant Science aslinks Meens Brass Mass Animals Automacy, Engineering Winter Science intertainment Social Science Instant, Science aslinks Meens Brass Mass Animals Automacy, Engineering Winter Science intertainment Social Science Inside Yabos istern Misser, Lee Taxe Animals Automacy, Engineering Winter Science intertainment Social Science Impage intertainment Social Scince Impage	WW, Software, Games Library	es. Dictionaties. Quotations	· Mobile phones, service plans and
Iducation Regional Beadcast Eve olige and University, K-12. Counties Regions, US Suites - intertainment Science adlarks, Mexin, Rome, Mass - iovernment Annals, Amonary, Expinenting- iovernment Social Science iovernment Social Science Inside Yahos iovernment Social Science Inside Yahos issenth Social Science Inside Yahos issenth Social Science Inside Yahos issenth Social Environment Regime Inside Participantic Science inflactions, Drugs, Emras - Exceller Environment Regime Social Science Inside Yahos inflacture Society & Culture Inside Participantic The Regime Inside Participantic The Regime inflacture Commerced by Commerced Inside Participantic The Regime Inside Participantic The Regime Level Yahote's Commerced by - Commerced by - Society - Span - Sweden - UK & Intend Intend			accessories
Arthurel Society & Culture Construct Regions US States Science	n Reg	ional	Broadcast Events
Intertainment Science asiliaria Moran Ilana Maar - Aninala Amonoury, Inginering - Inside Yahas iovernment Social Science Arthurelegy: Economic, Impraga Inside Yahas Social Science Arthurelegy: Economic, Impraga Inside Yahas Social Science Arthurelegy: Economic, Impraga Inside Yahas Social Science Arthurelegy: Economic, Impraga Inside Yahas Inside Ya	University, K-12. Cours	nes. Regions. US States	+ 4pm ET : POA: Western Opm
Inside Yahou Social Science Social Science Aritation Science Aritation Science Aritation Science Aritation Science Aritation Science Social Science Aritation Science Social Science Aritation Science Social Science Aritation Science Inside Yahou Social Science Aritation Science Inside Yahou Inside Inside Inside Yahou Inside Inside Science Inside Yahou Inside Yahou Inside Yahou Inside Yahou Inside Yahou Inside Science Inside Yahou Inside Inside Science Inside Yahou Inside Science Inside Yahou Inside Yahou Inside Yahou Inside Yahou Inside Science Inside Yahou Inside Science Inside Yahou Inside Science Inside Yahou Inside Science Inside Science Inside Yahou Inside Science Inside	Cole		· blink-182 - Artist of the month
iovernment iovernment Social Science Archaelegy, Economics, Languages isoking, Lee, Taxon Restore, Mittage, Lee, Taxon Restore, Mittage, Lee, Taxon Restore, Mittage, Lee, Social Science Social Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Scie	interns Deve	nce	DOM.
Sovernment Social Science Antherology: Economics, Languages Bealth Society & Culture Industre, Densier, Drugs, Farris, People, Environment, Religion Common Structure, Provide Environment, Religion Common Structure, Struc	STRUCT FOR THE ADDRESS ADDRESS ADDRESS	er verstent, rebestert-	Inside Yahoo!
Indexens Mainery, Lee, Taxon - Archaeology: Economics Languages - Indexens Danaes, Denga, Energia - Brogis Environment Religion - Indexens Danaes, Denga, Elevisa - Brogis Environment Religion - Indexens - Version - Start Sec UK & Information - Start Sec	nent Soc	al Science	- 11 Games - backgammon, eaches,
Easilih Society & Culture Compared by Compared by Compared Society & Society & Culture Society & Society & Culture Society & Society & Culture Society & Society & Society Society Society & Society	Mary Les. Taxes - Arths	ology, Economics, Languages-	hearts, chess, pinochle
Isabih Society & Culture Infini Danier, Drugo, Einsis, - Propie Environment Religen COMPAGE Local Values's Local Values's Local Values's	- the set of the set	A State of State of State	- Il Mevies - Scary Movie 2. Kinn
Indeiner, Disensons, Druge, Ettenson - Propie Environment, Religion - (COMPAGE) Local Yahoots	Soc	ety & Culture	of the Dragen, Cats and Degs
COMPAG · Vittatas - post year	aseases, Drugs, Eitness - Proph	Environment, Religion	- HEAT PROVIDE CARACO BALEDAR -
Local Yahools Zorope: Demank - France - Gemany - Indy - Nerway - Span - Sweden - UK & Intend	powwwed by	7	W Photos - most some name men
Local Yaboots Execute: Damain - France - Gename - Indy - Nervay - Span - Sweden - UK & Infand	COMPA	2)	and her her her her her her
Local Yahoots Europe - Demark - France - Genary - Infy - Nervay - Spain - Sweden - UK & Infand			
Europe : Denmark - France - Germany - Italy - Norway - Spain - Sweden - UK & Iteland	Yaloots .		
	· Denmark - Enance - Germany -	haly - Nervay - Span - Sweden -	UK & Ireland
Arte Pecific : Ania - Australia & NZ - China - HK - India - Japan - Kensa - Sangapara - Tauran	leepte : Ania - Australia & NZ - 9	hina - 25K - India - Japan - Kersa -	Sugapore - Lainat
dwartout : Argentina - Buzal - Canada - Chinese - Mesica - Spanish	our : Argentina - Brazil - Canada	Chinese - Mesica - Spanish	
U.S. Caner : Adanta - Beston - Chicago - Dallos FW - LA - NYC - SF Bay - Wath DC - mare-	aner: Adanta - Beston - Chicago	- Dates FW - LA - NYC - SE Bay	- Wath DC - mareau
Mare Value's	Valuela		

Make Taheol your home page

Small Buchmar: Biz Madeeplace - Domain Registration - Small Biz Center - Store Building - Web Hasting Access Fashool star: Pagers, PDAs, Web-enabled Phones and Voice (1.400-Ma-Yahoo)

Han to Suggest a Site - Company Info - Copyright Policy - Terms of Service - Contributors - Juba - Adventicing

Copyright @ 2001 Yahoo! Inc. All rights reserved. Privacy Policy
The Web , circa 1998





Match keywords, language (*information retrieval*) Explore directory





ew' create your own webcan	Click to deveload Yaboal Messenger	free from anywhere	
11 Shopping Depts: Bool actions Autor Chandleds St Carrers Chat Chills ScoCate 1 Addt Bools Briefrane Calm	ks. CDs. Computers, DVDs. Stores : hopping Travel Yellow Egs. Maps. Media 5 Greetings: Mill: Meuters: Messenger dar. My, Yahoo! FarDirect Fon Gauss	Gap. Clinique, Coach and more Eliance Quotes Marin Sparin Weath Mathe - Innionals - Fengle Search - Phi Kida - Marina - Masin - Rathe - DV - more	
Yahoo! Auction	S - Bid, burr, or sell anything!	In the News	
egories Aques Camputers mas Electronica ita Specta Carda nic Realca Stamps Baseball Cards - McGrain, A.R.	Items - Xeea - Date Eanihardt - Goff Chilos - Scafes Vesterday - ultid - Centralbergar - Dekenan - Icacam - Johnson - Icacam	 U.S. rejects 11M rgy plane tab Source: Conde admits to sexual relationship with missing interm Atterney Barry Levin found dead Dale Earthouth IV, with Papet 400 Withbedon - Tour de France 	
ts & Humanities	News & Media	Harkotplace	
unie, motography- siness & Economy , timmo, Shepping John- imputers & Internet met. W.W., Software, Ganca-	Recreation & Sports Spans Tanual Auras Cardeon Reference Linears Dictionains Containna-	- seret" Enney, sheps London, Epinoda 4 - spansared by Pepni 11 Disson - becomes part of Value of Shapping 11 Controls, of a job, pest your revenue Mobile planen, service plana and accessions	
ucation	Regional	Broadcast Events	
ege and Unit mility, K-12	Countries, negrand, U2 States	+ 4pm ET : POA: Western Opm	
		· Mink-112 - Artist of the month	

TATION

6

Shop A Connect Persons

24000

Line

212

Ed Col

Eler

Time-consuming; Not open-ended

Guide: Annue - Banz Belles - Carento - Houlbo - Lorney - Databona - Pars - Real Zarata - Tahashapani Bouwashwawi - Antanlogy - Reneducat - Events - Gauss - Maries - Massi - Radio - Takisa - Ty - more Powers - Banking - Bill Dy - Insurance - Lorne - Lanes - Famine/Philin - more Local - Chandida - Events - Ledgura - Marga - Reningura - Yalko Parga - more News - Type Stores - Banaras - Longura - Marga - Reningura - Yalko Parga - more News - Type Stores - Banaras - Insurance - Lornes - House - Registra - Jachadegy - Wanther Powershowy - Banaras - Carlos - Equation - Insura - House - Manage Banada Social Storetow: Ba Mathemphore - Donain Registration - Annual Ba Carton - Store Building - Web Hasting Access Tables - Var. Parges - Education - Markashow - House - Manage Lincola

Make Tabeol your home page

Haw to Suggest a Site - Company Info - Copyright Policy - Terms of Service - Contabutors - Jaba - Adventising

Copyright @ 2001 Yahoo! Inc. All rights reserved. Privacy Policy

Enter PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

Computer Science Department, Stanford University, Stanford, CA 94305, USA sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy

use of the structure and produce much 1 text and hyperlink c

The PageRank Citation Ranking: Bringing Order to the Web

January 29, 1998

...

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively

Key Idea: Consider the citations of the website.





Innovation 1: What pages would a "random Web surfer" end up at?

Innovation 2: Not just own terms but what terms are used by citations?



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Innovation 1: What pages would a "random Web surfer" end up at?

Innovation 2: Not just own terms but what terms are used by citations?

View 1: Flow Model:

in-links (citations) as votes

but, citations from important pages should count more.

=> Use recursion to figure out if each page is important.

Innovation 1: What pages would a "random Web surfer" end up at?

Innovation 2: Not just own terms but what terms are used by citations?

View 1: Flow Model:



How to compute?

View 1: Flow Model:

A
$$r_{A}/1$$
 B
C $r_{C}/2$ D $r_{D} = r_{A}/1 + r_{B}/4 + r_{C}/2$

How to compute?

View 1: Flow Model:



How to compute?

View 1: Flow Model:



A System of Equations:

 $r_A = \frac{r_B}{2} + \frac{r_C}{1}$

How to compute?

View 1: Flow Model:

A System of Equations:



How to compute?



How to compute?



$$r_{A} = \frac{r_{B}}{2} + \frac{r_{C}}{1}$$

$$r_{B} = \frac{r_{A}}{3} + \frac{r_{D}}{2}$$

$$r_{C} = \frac{r_{A}}{3} + \frac{r_{D}}{2}$$

$$r_{D} = \frac{r_{A}}{3} + \frac{r_{B}}{2}$$



$$1 = r_A + r_B + r_C + r_D$$



to \ from	Α	B	С	D
A	0	1/2	1	0
В	1/3	0	0	1/2
С	1/3	0	0	1/2
D	1/3	1/2	0	0

Transition Matrix, M

Innovation: What pages would a "random Web surfer" end up at? To start: N=4 nodes, so $r = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$

View 2: Matrix Formulation



$$1 = r_A + r_B + r_C + r_D$$



to \ from	Α	В	С	D
A	0	1/2	1	0
В	1/3	0	0	1/2
С	1/3	0	0	1/2
D	1/3	1/2	0	0

Transition Matrix, M

Innovation: What pages would a "random Web surfer" end up at?

To start: N=4 nodes, so $r = [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$

after 1st iteration: $M \cdot r = [3/8, 5/24, 5/24, 5/24]$ after 2nd iteration: $M(M \cdot r) = M^2 \cdot r = [15/48, 11/48, ...]$

View 2: Matrix Formulation

 $1 = r_A + r_B + r_C + r_D$



to \ from	Α	B	С	D
A	0	1/2	1	0
В	1/3	0	0	1/2
С	1/3	0	0	1/2
D	1/3	1/2	0	0

Transition Matrix, M



"Transition Matrix", M



"Transition Matrix", M

As err_norm gets smaller we are moving toward: $r = M \cdot r$

View 3: Eigenvectors:

Power iteration algorithm

As err_norm gets smaller we are moving toward: $r = M \cdot r$



As err_norm gets smaller we are moving toward: $r = M \cdot r$



View 4: Markov Process

Where is surfer at time t+1? $p(t+1) = M \cdot p(t)$

Suppose: p(t+1) = p(t), then p(t) is a *stationary distribution* of a *random walk*.

Thus, r is a stationary distribution. Probability of being at given node.

View 4: Markov Process

Where is surfer at time t+1? $p(t+1) = M \cdot p(t)$

Suppose: p(t+1) = p(t), then p(t) is a *stationary distribution* of a *random walk*.

Thus, r is a stationary distribution. Probability of being at given node.

aka 1st order Markov Process

- Rich probabilistic theory. One finding:
 - Stationary distributions have a unique distribution if:
 - No "dead-ends": a node can't propagate its rank
 - No *"spider traps"*: set of nodes with no way out.

View 4: Markov Process - Problems for vanilla PI



to \ from	Α	В	С	D
А	0	0	1	0
В	1/3	0	0	1
С	1/3	0	0	0
D	1/3	0	0	0

What would *r* converge to?

aka 1st order Markov Process

- Rich probabilistic theory. One finding:
 - Stationary distributions have a unique distribution if:
 - No "*dead-ends*": a node can't propagate its rank
 - No *"spider traps"*: set of nodes with no way out.

View 4: Markov Process - Problems for vanilla PI



to \ from	Α	В	С	D
А	0	0	1	0
В	1/3	0	0	1
С	1/3	0	0	0
D	1/3	1	0	0

What would *r* converge to?

aka 1st order Markov Process

- Rich probabilistic theory. One finding:
 - Stationary distributions have a unique distribution if:
 - No "dead-ends": a node can't propagate its rank
 - No "spider traps": set of nodes with no way out.

View 4: Markov Process - Problems for vanilla PI



to \ from	Α	В	С	D
А	0	0	1	0
В	1/3	0	0	1
С	1/3	0	0	0
D	1/3	1	0	0

What would *r* converge to?

aka 1st order Markov Process

- Rich probabilistic theory. One finding:
 - Stationary distributions have a unique distribution if:

same node doesn't repeat at regular intervals columns sum to 1 non-zero chance of going to any other node

The "Google" PageRank Formulation

- 1. Follow a random link (probability, $\beta = \sim .85$)
- 2. Teleport to a random node (probability, $1-\beta$)





The "Google" PageRank Formulation

- 1. Follow a random link (probability, $\beta = \sim.85$)
- 2. Teleport to a random node (probability, $1-\beta$)

	to \ from	Α	В	С	D
Y	А	0	0	1	0
	В	1⁄3	0	0	1
)	С	1⁄3	0	0	0
	D	1⁄3	1	0	0



The "Google" PageRank Formulation

- 1. Follow a random link (probability, $\beta = \sim.85$)
- 2. Teleport to a random node (probability, $1-\beta$)

	to \ from	А	В	С	D
Y	A	0	0+.15*1⁄4	1	0+.15*1⁄4
	В	1⁄3	0+.15*1⁄4	0	.85*1+.15*1⁄4
)	С	1⁄3	0+.15*1⁄4	0	0+.15*1⁄4
	D	1⁄3	.85*1 +.15* ¹ ⁄4	0	0+.15*1⁄4



The "Google" PageRank Formulation

- 1. Follow a random link (probability, $\beta = \sim.85$)
- 2. Teleport to a random node (probability, $1-\beta$)

to \ from	Α	В	С	D
А	0+.15*1⁄4	0+.15*1⁄4	85*1+. <mark>15</mark> *¼	0+.15*1⁄4
В	.85* ¹ ⁄ ₃ +.15* ¹ ⁄ ₄	0+.15*1⁄4	0+.15*1⁄4	.85*1+. <mark>15</mark> *¼
С	.85* ¹ ⁄ ₃ +.15* ¹ ⁄ ₄	0+.15*1⁄4	0+.15*1⁄4	0+.15*1⁄4
D	.85* ¹ ⁄ ₃ +.15* ¹ ⁄ ₄	.85*1+.15*1⁄4	0+.15*1⁄4	0+.15*1⁄4



The "Google" PageRank Formulation

- 1. Follow a random link (probability, $\beta = \sim.85$)
- 2. Teleport to a random node (probability, $1-\beta$)



to \ from	Α	В	С	D
A	0	0	1	0
В	1⁄3	0	0	1
С	1⁄3	0	0	0
D	1⁄3	0	0	0

The "Google" PageRank Formulation

- 1. Follow a random link (probability, $\beta = \sim.85$)
- 2. Teleport to a random node (probability, $1-\beta$)



to \ from	Α	В	С	D
Α	0	1⁄4	1	0
В	1⁄3	1⁄4	0	1
С	1⁄3	1⁄4	0	0
D	1/3	1⁄4	0	0

The "Google" PageRank Formulation

- 1. Follow a random link (probability, $\beta = \sim.85$)
- 2. Teleport to a random node (probability, $1-\beta$)



to \ from	Α	В	С	D
Α	0	.85*1⁄4+.15*1⁄4	1	0
B	1⁄3	.85*1⁄4+.15*1⁄4	0	1
С	1⁄3	.85*1⁄4+.15*1⁄4	0	0
D	1⁄3	.85*1⁄4+.15*1⁄4	0	0

The "Google" PageRank Formulation

Add teleportation: At each step, two choices

- 1. Follow a random link (probability, $\beta = \sim.85$)
- 2. Teleport to a random node (probability, $1-\beta$)

(Teleport from a dead-end has probability 1)



to \ from	Α	B	С	D
A	0+.15*1⁄4	1*1⁄4	85*1+. <mark>15</mark> *¼	0+.15*1⁄4
В	.85*1⁄3+.15*1⁄4	1*1⁄4	0+.15*1⁄4	.85*1+. <mark>15</mark> *¼
С	.85*½+.15*¼	1*1⁄4	0+.15*1⁄4	0+.15*1⁄4
D	.85*½+.15*¼	1*1⁄4	0+.15*1⁄4	0+.15*1⁄4

Teleportation, as Flow Model:

$$r_{j} = \sum_{i \to j} \beta \frac{r_{i}}{d_{i}} + (1 - \beta) \frac{1}{N}$$
(Brin and Page, 1998)



to \ from	Α	В	С	D
A	0+.15*1⁄4	1*1⁄4	85*1+. <mark>15</mark> *¼	0+.15*1⁄4
В	.85* ¹ ⁄ ₃ +.15* ¹ ⁄ ₄	1*1⁄4	0+.15*1⁄4	.85*1+. <mark>15</mark> *¼
С	.85* ¹ ⁄ ₃ +.15* ¹ ⁄ ₄	1*1⁄4	0+.15*1⁄4	0+.15*1⁄4
D	.85*1⁄3+.15*1⁄4	1*1⁄4	0+.15*1⁄4	0+.15*1⁄4



Teleportation, as Flow Model:

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{N}$$
(Brin and Page, 1998)

Teleportation, as Matrix Model: $M' = \beta M + (1 - \beta) \left[\frac{1}{N}\right]_{N \times N}$



to \ from	А	В	С	D
A	0+.15*1⁄4	1*1⁄4	85*1+. <mark>15</mark> *¼	0+.15*1⁄4
В	.85* ¹ ⁄ ₃ +.15* ¹ ⁄ ₄	1*1⁄4	0+.15*1⁄4	.85*1+. <mark>15</mark> *¼
С	.85* ¹ ⁄ ₃ +.15* ¹ ⁄ ₄	1*1⁄4	0+.15*1⁄4	0+.15*1⁄4
D	.85*1⁄3+.15*1⁄4	1*1⁄4	0+.15*1⁄4	0+.15*1⁄4
Goals: No "dead-ends" No "spider traps" Teleportation, as Flow Model:

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{N}$$
(Brin and Page, 1998)

Teleportation, as Matrix Model: $M' = \beta M + (1 - \beta) \left[\frac{1}{N}\right]_{N \times N}$

to \ from	A	В	С	D
A	0+.15*1⁄4	.85*1⁄4+.15*1⁄4	85*1+.15*1⁄4	0+.15*1⁄4
В	.85*1⁄3+.15*1⁄4	.85*1⁄4+.15*1⁄4	0+.15*1⁄4	.85*1+.15*¼
С	.85*1⁄3+.15*1⁄4	.85*1⁄4+.15*1⁄4	0+.15*1⁄4	0+.15*1⁄4
D	.85*1⁄3+.15*1⁄4	.85*1⁄4+.15*1⁄4	0+.15*1⁄4	0+.15*1⁄4



Goals: No "dead-ends" No "spider traps"

Teleportation, as Flow Model:

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{N}$$
(Brin and Page, 1998)

Teleportation, as Matrix Model: $M' = \beta M + (1 - \beta) \left[\frac{1}{N}\right]_{M'}$

Steps:

- 1. Compute M
- 2. Add 1/N to all dead-ends.
- 3. Convert *M* to *M*'
- 4. Run Power Iterations.

to \ from	Α	В	С	D
Α	0+.15*1⁄4	1*1⁄4	85*1+.15*¼	0+.15*1⁄4
В	.85*1⁄3+.15*1⁄4	1*1⁄4	0+.15*1⁄4	.85*1+.15*¼
С	.85*1⁄3+.15*1⁄4	1*1⁄4	0+.15*1⁄4	0+.15*1⁄4
D	.85*1⁄3+.15*1⁄4	1*1⁄4	0+.15*1⁄4	0+.15*1⁄4